

Kernel Spectral Clustering

Ilaria Giulini*
 ilaria.giulini@me.com

June 22, 2016

Abstract

We investigate the question of studying spectral clustering in a Hilbert space where the set of points to cluster are drawn i.i.d. according to an unknown probability distribution whose support is a union of compact connected components. We modify the algorithm proposed by Ng, Jordan and Weiss in [15] in order to propose a new algorithm that automatically estimates the number of clusters and we characterize the convergence of this new algorithm in terms of convergence of Gram operators. We also give a hint of how this approach may lead to learn transformation-invariant representations in the context of image classification.

Keywords. Spectral clustering, Reproducing kernel Hilbert space, Markov chain, Gram operator

1 Introduction

Clustering is the task of grouping a set of objects into classes, called *clusters*, in such a way that objects in the same group are more similar to each other than to those in other groups. Spectral clustering algorithms are efficiently used as an alternative to classical clustering algorithms, such as k -means, in particular in the case of not linearly separable data sets. To perform clustering, these methods use the spectrum of some data-dependent matrices: the affinity matrix [7], or the Laplacian matrix [8]. Many different versions of spectral clustering algorithms can be found in the literature (see [12]) and

*The results presented in this paper were obtained while the author was preparing her PhD under the supervision of Olivier Catoni at the Département de Mathématiques et Applications, École Normale Supérieure, Paris, with the financial support of the Région Île de France.

one of the most successful algorithm has been proposed by Ng, Jordan and Weiss in [15]. Given a set of points to cluster into c classes and denoting by A the affinity matrix and by D the diagonal matrix whose i -th entry is the sum of the elements on the i -th row of A , their algorithm uses the c largest eigenvectors of the Laplacian matrix $D^{-1/2}AD^{-1/2}$ simultaneously. More precisely, the data set is first embedded in a c -dimensional space in which clusters are more evident and then points are separated using k -means, or any other standard algorithm.

Other algorithms use different renormalizations of the affinity matrix. For example, Shi and Malik in [16] use the second smallest eigenvector of the unnormalized Laplacian matrix $D - A$ to separate the data set into two groups and use the algorithm recursively to get more than two classes, whereas Meila and Shi in [14] use the first c largest eigenvectors of the stochastic matrix $D^{-1}A$ (that has the same eigenvectors as the normalized Laplacian matrix $I - D^{-1}A$) to compute a partition in c classes. These algorithms treat the question as a graph partitioning problem and they are based on the so called normalized cut criterion. Different cost functions can be found in the literature (see [6], [11]) and more generally the study of Laplacian matrices has been carry out also in different contexts, as for semi-supervised learning [5] and manifold learning [2].

We consider the setting of performing spectral clustering in a Hilbert space. This general framework includes the analysis of both data sets in a functional space and samples embedded in a reproducing kernel Hilbert space. The latter is the case of kernel methods that use the kernel trick to embed the data set into a reproducing kernel Hilbert space in order to get a new representation that simplifies the geometry of the classes. Our point of departure is the algorithm proposed by Ng, Jordan and Weiss [15] but interpreted in a infinite-dimensional setting, so that we view the matrices described above as empirical versions of some underlying integral operators. We assume that the points to cluster are drawn according to an unknown probability distribution whose support is a union of compact connected components (see [17] for consistency of clustering algorithms). Our idea is to view spectral clustering as a change of representation in a reproducing kernel Hilbert space, induced by a change of kernel, and to propose a new algorithm that automatically estimates the number of clusters. Usually the number of clusters is assumed to be know in advance (see [14], [15]) or it is linked to the presence of a sufficient large gap in the spectrum of the Laplacian matrix. To achieve our aim we replace the projection on the space spanned by the largest eigenvectors of the Laplacian matrix proposed by Ng, Jordan

and Weiss with a suitable power of the Laplacian operator. Indeed, such an iteration performs some kind of soft truncation of the eigenvalues and hence leads to a natural dimensionality reduction. This iteration is related to the computation of the marginal distribution at time t of some Markov chains with exponential rare transitions, as suggested by [4] in the case where the unknown probability distribution has a finite support. We conjecture (and this will be the subject of a future work) that the same kind of behavior holds for more general supports and we hint that spectral clustering, coupled with some preliminary change of representation in a reproducing kernel Hilbert space, can be a winning tool to bring down the representation of classes to a low-dimensional space and may lead to a generic and rather radical alternative. This suggests that the kernel trick, introduced to better separate classes in the supervised learning framework addressed by support vector machines (SVMs), is also feasible in the unsupervised context, where we do not separate classes using hyperplanes in the feature space but instead we use spectral clustering to perform the classification. We also suggest with an example how this approach may lead to learn transformation-invariant representations of a same pattern.

Developing a convincing toolbox for unsupervised invariant-shape analysis is beyond the scope of this study and it will be carried on elsewhere. However we observe that the pattern transformations we would like to take into account in image analysis are numerous and not easy to formalize: they may come from some set of transformations such as translations, rotations or scaling or they may come from the conditions in which the images have been taken, for example, changes in the perspective or in illumination, partial occlusions, object deformations, etc. Making a representation invariant with respect to a set of transformations is a challenging task even in the simpler case of translations. Indeed, it is sufficient to observe that the set of functions obtained by translating a single pattern in various directions typically spans a vector space of high dimension, meaning that the shapes (here the functions) that we would like to put in the same category do not even live in a common low-dimensional subspace. A possible approach is to study representations that leave invariant some group of transformations, for instance, the Fourier transform that has translation invariant properties, since its modulus is translation invariant. However it is unstable to small deformations at high frequencies. Wavelet transforms provide a workaround. Scattering representations proposed by Mallat [13] compute translation-invariant representations by cascading wavelet transforms and modulus pooling operators. They bring improvements for audio [1] and for image [3] classification. This kind of careful mathematical study has to be repeated for any kind of

transformations. Instead of deriving the representation of a pattern from a mathematical study, the idea here is to learn the representation itself from examples of patterns that sample in a sufficiently dense way the orbits of the set of transformations at stake.

The paper is organized as follows. In Section 2 we introduce the ideal version of our algorithm that uses the underlying unknown probability distribution and we provide an interpretation of the clustering effect in terms of Markov chains. In Section 3 we introduce an empirical version of the algorithm and we provide some convergence results, based on the convergence of some Gram operators. Finally experiments are shown in Section 4.

2 The Ideal Algorithm

Let \mathcal{X} be a compact subset of some separable Hilbert space endowed with the (unknown) probability distribution P on \mathcal{X} and assume that the support of P is made of several compact connected components. Let $A : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a symmetric positive semi-definite kernel, normalized in such a way that $A(x, x) = 1$, for any x in the state space \mathcal{X} , and let us define the symmetric positive semi-definite kernel

$$K(x, y) = A(x, y)^2.$$

This kernel plays the same role of the affinity matrix, while the Laplacian matrix is now replaced by the integral operator with kernel

$$\bar{K}(x, y) = \mu(x)^{-1/2} K(x, y) \mu(y)^{-1/2}$$

where the function

$$\mu(x) = \int K(x, z) dP(z) \tag{1}$$

is the analogous of the diagonal matrix D . According to the Ng, Jordan and Weiss algorithm we consider more specifically the case where

$$K(x, y) = K_\beta(x, y) = \exp(-\beta \|x - y\|^2), \quad \beta > 0.$$

2.1 The Algorithm

The ideal algorithm (that uses the unknown probability distribution P) goes as follows. For any $x, y \in \mathcal{X}$,

1. Form the kernel

$$\bar{K}(x, y) = \mu(x)^{-1/2} K(x, y) \mu(y)^{-1/2}$$

where $\mu(x) = \int K(x, z) dP(z)$

2. Construct the new (iterated) kernel

$$\bar{K}_m(x, y) = \int \bar{K}(y, z_1) \bar{K}(z_1, z_2) \dots \bar{K}(z_{m-1}, x) dP^{\otimes(m-1)}(z_1, \dots, z_{m-1}),$$

where $m > 0$ is a free parameter

3. Make a last change of kernel (normalizing the kernel \bar{K}_m)

$$K_m(x, y) = \bar{K}_{2m}(x, x)^{-1/2} \bar{K}_{2m}(x, y) \bar{K}_{2m}(y, y)^{-1/2}$$

4. Cluster points according to the new representation defined by the kernel K_m .

As it is suggested in the next section, the representation induced by the kernel K_m makes the subsequent classification an easy task.

2.2 The Clustering Effect

The main idea is that, in the Hilbert space defined by the kernel K_m , clusters are concentrated around orthogonal unit vectors, forming the vertices of a regular simplex. To give an intuition of this clustering effect, we provide a Markov chain analysis of the algorithm, according to some ideas suggested by [4].

Assume that the scale parameter β in the Gaussian kernel K is large enough. For the choice of β we refer to Section 2.3.

Define the kernel

$$M(x, y) = \mu(x)^{-1} K(x, y)$$

where μ is defined in equation (1) and consider the corresponding integral operator

$$\mathbf{M}(f) : x \mapsto \int M(x, z) f(z) dP(z).$$

Observe that \mathbf{M} is the transition operator of a Markov chain $(Z_k)_{k \in \mathbb{N}}$ on \mathcal{X} and that

$$M(x, y) = \frac{dP_{Z_1|Z_0=x}}{dP}(y).$$

So the Markov chain related to the operator \mathbf{M} has a small probability to jump from one component to another one. As already said, from [4] it can be deduced that, when the support of \mathbf{P} is finite, for suitable values of m , depending on β as $\exp(\beta T^2)$, the measure

$$\mathbf{M}_x^m : f \mapsto \mathbf{M}^m(f)(x)$$

is close (for large enough values of β) to be supported by a cycle of depth larger than H of the state space. The cycle decomposition of the state space of a Markov chain with exponential transitions is some discrete counterpart of the decomposition into connected components. More precisely, the measure \mathbf{M}_x^m is close to the invariant measure of the operator \mathbf{M} restricted to the connected component which x belongs to. As a result, the function $x \mapsto \mathbf{M}^m(f)(x)$ is approximately constant on each connected component. Thus, for reasons that we will not try to prove here, but that are confirmed by experiments, we expect the same kind of behaviour in the general setting. To rewrite this conjecture in terms of the kernel K_m we introduce the following proposition.

Proposition 1. *Let \mathbf{Q} be the invariant measure of the Markov transition operator \mathbf{M} and define*

$$\mathfrak{R}_x = \mu(x)^{1/2} \frac{d\mathbf{M}_x^m}{d\mathbf{Q}} \in L_{\mathbf{Q}}^2, \quad x \in \mathcal{X}.$$

It holds that

$$K_m(x, y) = \left\langle \frac{\mathfrak{R}_x}{\|\mathfrak{R}_x\|_{L_{\mathbf{Q}}^2}}, \frac{\mathfrak{R}_y}{\|\mathfrak{R}_y\|_{L_{\mathbf{Q}}^2}} \right\rangle_{L_{\mathbf{Q}}^2}.$$

Proof. Using the fact that $\bar{K} = \bar{K}_1$ and that

$$\bar{K}_m(x, y) = \int \bar{K}(x, z) \bar{K}_{m-1}(z, y) d\mathbf{P}(z), \quad (2)$$

by induction we get

$$\mathbf{M}^m(f)(x) = \mu(x)^{-1/2} \int \bar{K}_m(x, z) \mu(z)^{1/2} f(z) d\mathbf{P}(z).$$

Thus the measure \mathbf{M}_x^m has density

$$\frac{d\mathbf{M}_x^m}{d\mathbf{P}}(y) = \mu(x)^{-1/2} \bar{K}_m(x, y) \mu(y)^{1/2}.$$

Moreover, since

$$\int \mu(x)M(x, y)f(y) \, dP(x)dP(y) = \int K(x, y)f(y) \, dP(x)dP(y) = \int \mu(y)f(y) \, dP(y),$$

the invariant measure Q has a density with respect to P equal to $\frac{dQ}{dP} = \mu$.

As a consequence

$$\frac{d\mathbf{M}_x^m}{dQ}(y) = \mu(x)^{-1/2}\overline{K}_m(x, y) \mu(y)^{-1/2}.$$

According to equation (2) and recalling the definition of K_m we conclude the proof. ■ □

With these definitions, our conjecture can be formulated as

$$\lim_{\beta \rightarrow \infty} K_{\exp(\beta T^2)}(x, y) = \sum_{C \in \mathcal{C}_T} \mathbb{1}(\{x, y\} \subset C),$$

where \mathcal{C}_T denotes the connected components of the graph $\{x, y \in \text{supp}(P) \mid \|y - x\| < T\}$.

Remark that, as we assumed that the support of P is a union of compact topological connected components, taking T to be less than the minimum distance between two topological components of $\text{supp}(P)$, we obtain that \mathcal{C}_T coincides with the topological components of $\text{supp}(P)$. As a consequence, for suitable values of the scale parameter β and of the number of iterations m , the kernel $K_m(x, y)$ is close to zero (or equivalently the supports of the probability measures \mathbf{M}_x^m and \mathbf{M}_y^m are almost disjoint) if x and y belong to two different clusters, whereas it is close to one (or equivalently the supports of \mathbf{M}_x^m and \mathbf{M}_y^m are almost the same) when x and y belong to the same cluster. Moreover since the kernel $K_m(x, y)$ is the cosine of the angle formed by the two vectors representing x and y , according to the conjecture, it is either close to zero or close to one, showing that, in the Hilbert space defined by K_m , clusters are concentrated around orthogonal unit vectors.

2.3 Choice of the Scale Parameter

We conclude this section with some remarks on the choice of the scale parameter β . In the case where the influence kernel K is of the form

$$K(x, y) = K_\beta(x, y) = \exp(-\beta\|x - y\|^2),$$

we propose to choose β as the solution of the equation

$$F(\beta) := \int K_\beta(x, y)^2 \, dP(x)dP(y) = h$$

where h is a suitable parameter which measures the probability that two independent points drawn according to the probability P are close to each other. Introducing the Gram operator $L_\beta : L_P^2 \rightarrow L_P^2$ defined by the kernel K_β as

$$L_\beta(f)(x) = \int K_\beta(x, y) f(y) dP(y), \quad x \in \mathcal{X},$$

the parameter h is equal to the square of the Hilbert-Schmidt norm of L_β . Observe that L_β has a discrete spectrum $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$ that satisfies, according to the Mercer theorem,

$$\begin{aligned} \sum_{i=1}^{+\infty} \lambda_i &= \int K_\beta(x, x) dP(x) = 1, \\ \sum_{i=1}^{+\infty} \lambda_i^2 &= F(\beta) \leq 1 \end{aligned}$$

since $K_\beta(x, x) = 1$. We also observe that

$$\lim_{\beta \rightarrow 0} F(\beta) = 1$$

which implies that for β going to zero, λ_1 goes to one whereas all the other eigenvalues λ_i , for $i \geq 2$, go to zero. Moreover,

$$\lim_{\beta \rightarrow +\infty} F(\beta) = 0,$$

so that when β grows, the eigenvalues are spread more and more widely. Therefore, $F(\beta)$ governs the spread of the eigenvalues of L_β . For these reasons, the value of the parameter $h = F(\beta)$ controls the effective dimension of the representation of the distribution of points P in the reproducing kernel Hilbert space defined by K_β . Experiments show that this effective dimension has to be pretty large in order to have a proper clustering effect, meaning that we will impose a small value of the parameter h .

Note that in the experiments we do not have access to the function $F(\beta)$ but we have to consider its empirical version

$$\frac{1}{n(n-1)} \sum_{1 \leq i < j \leq n} K_\beta(X_i, X_j)^2$$

where X_1, \dots, X_n is an i.i.d. sample drawn from the probability distribution P .

3 The Empirical Algorithm

In order to provide an empirical version of the ideal algorithm proposed in Section 2, we connect the previous kernels with some Gram operators. Note that by *empirical algorithm* we mean an algorithm based on the sample distribution instead of on the unknown probability distribution P .

From now on, given a Hilbert space \mathcal{K} and a probability distribution \mathcal{P} on \mathcal{K} , the Gram operator $\mathfrak{G} : \mathcal{K} \rightarrow \mathcal{K}$ is defined as

$$\mathfrak{G}u = \mathbb{E}_{Y \sim \mathcal{P}}[\langle u, Y \rangle_{\mathcal{K}} Y].$$

The next proposition links the function μ defined in equation (1) to the Gram operator defined in the reproducing kernel Hilbert space induced by the kernel

$$A(x, y) = K(x, y)^{1/2} = \exp\left(-\beta\|x - y\|^2/2\right).$$

Proposition 2. *Let \mathcal{H}_A be the reproducing kernel Hilbert space defined by A and $\phi_A : \mathcal{X} \rightarrow \mathcal{H}_A$ the corresponding feature map. Denote by $\mathcal{G}_A : \mathcal{H}_A \rightarrow \mathcal{H}_A$ the Gram operator*

$$\mathcal{G}_A v = \int \langle v, \phi_A(z) \rangle_{\mathcal{H}_A} \phi_A(z) dP(z).$$

Then

$$\mu(x) = \int \langle \phi_A(x), \phi_A(z) \rangle_{\mathcal{H}_A}^2 dP(z) = \langle \mathcal{G}_A \phi_A(x), \phi_A(x) \rangle_{\mathcal{H}_A}.$$

Proof. The result follows recalling that by the Moore-Aronszajn theorem

$$A(x, y) = \langle \phi_A(x), \phi_A(y) \rangle_{\mathcal{H}_A}.$$

■

□

A similar result relates the iterated kernel \overline{K}_m to the Gram operator defined in the reproducing kernel Hilbert space induced by the Gaussian kernel K .

Proposition 3. *Let \mathcal{H} be the reproducing kernel Hilbert space defined by K and $\phi_K : \mathcal{X} \rightarrow \mathcal{H}$ the corresponding feature map. Define*

$$\phi_{\overline{K}}(x) = \mu(x)^{-1/2} \phi_K(x) \tag{3}$$

and consider the Gram operator $\mathcal{G} : \mathcal{H} \rightarrow \mathcal{H}$

$$\mathcal{G}v = \int \langle v, \phi_{\overline{K}}(z) \rangle_{\mathcal{H}} \phi_{\overline{K}}(z) dP(z). \tag{4}$$

Then, for any $m > 0$,

$$\overline{K}_{2m}(x, y) = \langle \mathcal{G}^{m-1/2} \phi_{\overline{K}}(x), \mathcal{G}^{m-1/2} \phi_{\overline{K}}(y) \rangle_{\mathcal{H}}.$$

Proof. According to the Moore-Aronszajn theorem

$$K(x, y) = \langle \phi_K(y), \phi_K(x) \rangle_{\mathcal{H}}$$

and thus, by definition, the kernel $\overline{K}(x, y) = \mu(x)^{-1/2} K(x, y) \mu(y)^{-1/2}$ can be written as

$$\overline{K}(x, y) = \langle \mu(x)^{-1/2} \phi_K(x), \mu(y)^{-1/2} \phi_K(y) \rangle_{\mathcal{H}} = \langle \phi_{\overline{K}}(y), \phi_{\overline{K}}(x) \rangle_{\mathcal{H}}.$$

Using the fact that $\overline{K} = \overline{K}_1$ and that

$$\overline{K}_m(x, y) = \int \overline{K}(x, z) \overline{K}_{m-1}(z, y) dP(z),$$

by induction we get that

$$\overline{K}_m(x, y) = \langle \mathcal{G}^{m-1} \phi_{\overline{K}}(x), \phi_{\overline{K}}(y) \rangle_{\mathcal{H}}.$$

Since \mathcal{G} is a positive symmetric operator we conclude the proof. ■

□

As a consequence, the representation of $x \in \mathcal{X}$ defined by the renormalized kernel K_m is isometric to the representation $\|\mathcal{R}_x\|_{\mathcal{H}}^{-1} \mathcal{R}_x \in \mathcal{H}$, where

$$\mathcal{R}_x = \mathcal{R}_x(m) = \mathcal{G}^{m-1/2} \phi_{\overline{K}}(x) \in \mathbf{Im}(\mathcal{G}) \subset \mathcal{H}.$$

Remark 4. The representation of x in the kernel space defined by K_m is also isometric to the representation $\|\mathbf{R}_x\|^{-1} \mathbf{R}_x \in L_{\mathbb{P}}^2$, where $\mathbf{R}_x \in L_{\mathbb{P}}^2$ is defined as

$$\mathbf{R}_x(z) = \overline{K}_m(x, z) = \langle \mathcal{G}^{m-1} \phi_{\overline{K}}(x), \phi_{\overline{K}}(z) \rangle_{\mathcal{H}}. \quad (5)$$

This is a consequence of the fact that

$$\langle \mathbf{R}_x, \mathbf{R}_y \rangle_{L_{\mathbb{P}}^2} = \int \overline{K}_m(x, z) \overline{K}_m(z, y) dP(z) = \overline{K}_{2m}(x, y).$$

3.1 An Intermediate Step

As already said, by the Moore-Aronszajn theorem, the Gaussian kernel K defines a reproducing kernel Hilbert space \mathcal{H} and a feature map $\phi_K : \mathcal{X} \rightarrow \mathcal{H}$ such that

$$K(x, y) = \langle \phi_K(y), \phi_K(x) \rangle_{\mathcal{H}}.$$

We introduce an intermediate version of the algorithm, that uses the feature map ϕ_K . Since this feature map is not explicit, we will afterward translate this description into an algorithm that manipulates only scalar products and not implicit feature maps. This intermediate step is useful to provide the convergence results presented in Section 3.3.

Let X_1, \dots, X_n be the set of points to cluster, that we assume to be drawn i.i.d. according to P . The algorithm goes as follows.

1. Form the kernel

$$\widehat{K}(x, y) = \langle \phi_{\widehat{K}}(x), \phi_{\widehat{K}}(y) \rangle_{\mathcal{H}}$$

where

$$\phi_{\widehat{K}}(x) = \widehat{\mu}(x)^{-1/2} \phi_K(x) \quad \text{and} \quad \widehat{\mu}(x) = \frac{1}{n} \sum_{i=1}^n K(x, X_i)$$

2. Construct the kernel

$$\widehat{K}_m(x, y) = \langle \widehat{\mathcal{Q}}^{\frac{m-1}{2}} \phi_{\widehat{K}}(x), \widehat{\mathcal{Q}}^{\frac{m-1}{2}} \phi_{\widehat{K}}(y) \rangle_{\mathcal{H}} \quad (6)$$

where

$$\widehat{\mathcal{Q}}v = \frac{1}{n} \sum_{i=1}^n \widehat{\mu}(X_i)^{-1} \langle v, \phi_K(X_i) \rangle_{\mathcal{H}} \phi_K(X_i).$$

The definition of $\widehat{\mathcal{Q}}$ is justified in Remark 5.

3. Make a last change of kernel and consider

$$H_m(x, y) = \widehat{K}_{2m}(x, x)^{-1/2} \widehat{K}_{2m}(x, y) \widehat{K}_{2m}(y, y)^{-1/2}$$

4. Cluster points according to this new representation, by thresholding the distance between points.

Remark 5. *The construction of the estimator $\hat{\mathcal{Q}}$ follows from a two-steps estimate of the Gram operator \mathcal{G} defined in equation (4). Indeed, according to the definition of the feature map $\phi_{\overline{K}}$, the Gram operator \mathcal{G} rewrites as*

$$\mathcal{G}v = \int \mu(z)^{-1} \langle v, \phi_K(z) \rangle_{\mathcal{H}} \phi_K(z) \, dP(z).$$

Thus first we replace the function μ with its estimator $\hat{\mu}$ and then we replace the unknown distribution P with the sample distribution.

3.2 Implementation of the Algorithm

We now describe the algorithm used in the experiments. As in the previous section, let X_1, \dots, X_n be the set of points to cluster, drawn i.i.d. according to P .

1. Construct for $i, j = 1, \dots, n$

$$K_{ij} = \frac{1}{n} \exp(-\beta \|X_i - X_j\|^2),$$

where the parameter β is chosen, according to Section 2.3, as the solution of

$$\frac{1}{n(n-1)} \sum_{\substack{i,j \\ i \neq j}} \exp(-2\beta \|X_i - X_j\|^2) \simeq 0.005$$

2. Define the diagonal matrix $D = \text{diag}(D_1, \dots, D_n)$ where

$$D_i = \max \left\{ \frac{1}{n} \sum_{j=1}^n K_{ij}, \sigma \right\} \quad \text{with } \sigma = 0.001$$

3. Form the matrix

$$M = D^{-1/2} K D^{-1/2}$$

4. Consider

$$C_{ij} = (M^m)_{ii}^{-1/2} (M^m)_{ij} (M^m)_{jj}^{-1/2}, \quad i, j = 1, \dots, n \quad (7)$$

The choice of the number of iterations m is done automatically and it is described in Remark 6.

5. Cluster points according to the representation induced by C , as detailed in Remark 7.

Note that in some configurations the above algorithm obviously yields an unstable classification, but in practice, when the classes are clearly separated from each other, this simple scheme is successful.

Remark 6. *To choose automatically the number of iterations, we have to fix a maximal number of clusters. This means that we have to provide an upper bound, that we denote by p , on the number of classes we expect to have. However, as it can be seen in the simulations, this choice of p is robust, meaning that p can be harmlessly overestimated. Thus, denoting by $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_n \geq 0$ the eigenvalues of M , we choose the number m of iterations by solving*

$$\left(\frac{\hat{\lambda}_p}{\hat{\lambda}_1}\right)^m \simeq \zeta$$

where $\zeta > 0$ is a given small parameter. The choice of ζ is also robust and $1/100$ is a reasonable value for it.

Remark 7. *Our (greedy) classification algorithm consists in taking an index $i_1 \in \{1, \dots, n\}$ at random, in forming the corresponding class*

$$\hat{C}_{i_1} = \left\{ j \mid C_{i_1 j} \geq s \right\},$$

where s is a threshold that we take equal to 0.1 in all our experiments, and then in starting again with the remaining indices. In such a way we construct a sequence of indices i_1, \dots, i_c , where, as a consequence, the number of clusters c is automatically estimated. The k -th class is hence defined as

$$\tilde{C}_k = \hat{C}_{i_k} \setminus \bigcup_{\ell < k} \hat{C}_{i_\ell}.$$

3.3 Convergence Results

We introduce some results on the accuracy of the estimation of the ideal algorithm through the empirical one. In particular we compare the ideal iterated kernel \bar{K}_{2m} with its estimator \widehat{K}_{2m} defined in equation (6).

Proposition 8. *Define*

$$\chi(x) = \left(\frac{\mu(x)}{\widehat{\mu}(x)} \right)^{1/2}. \quad (8)$$

For any $x, y \in \text{supp}(\mathbf{P})$, $m > 0$,

$$\begin{aligned} |\widehat{K}_{2m}(x, y) - \overline{K}_{2m}(x, y)| &\leq \frac{\max\{1, \|\chi\|_\infty\}^2}{\mu(x)^{1/2}\mu(y)^{1/2}} \left((2m-1)\|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty \left(1 + \|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty\right)^{2m-2} \right. \\ &\quad \left. + 2\|\chi - 1\|_\infty \right) \end{aligned}$$

and

$$\begin{aligned} \left\| \widehat{K}_{2m}(x, \cdot) - \overline{K}_{2m}(x, \cdot) \right\|_{L^2_{\mathbf{P}}} &\leq \frac{\max\{1, \|\chi\|_\infty\}^2}{\mu(x)^{1/2}} \left((2m-1)\|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty \left(1 + \|\widehat{\mathcal{Q}} - \mathcal{G}\|_\infty\right)^{2m-2} \right. \\ &\quad \left. + 2\|\chi - 1\|_\infty \right). \end{aligned}$$

For the proof we refer to Section 5.1.

Other convergence results are related to the description of the kernel K_m in terms of the representation $\mathbf{R}_x \in L^2_{\mathbf{P}}$ introduced in equation (5). More precisely, we recall that according to Proposition 3

$$K_m(x, y) = \left\langle \|\mathbf{R}_x\|^{-1} \mathbf{R}_x, \|\mathbf{R}_y\|^{-1} \mathbf{R}_y \right\rangle_{L^2_{\mathbf{P}}},$$

where $\mathbf{R}_x(z) = \langle \mathcal{G}^{m-1} \phi_{\overline{K}}(x), \phi_{\overline{K}}(z) \rangle_{\mathcal{H}}$ and the feature map $\phi_{\overline{K}}$ is defined in equation (3). Our estimated representation of x in $L^2_{\mathbf{P}}$ is $\widehat{\mathbf{N}}(x)^{-1} \widehat{\mathbf{R}}_x$, where

$$\widehat{\mathbf{R}}_x(y) = \langle \widehat{\mathcal{Q}}^{m-1} \phi_{\widehat{K}}(x), \phi_{\widehat{K}}(y) \rangle_{\mathcal{H}} \quad \text{and} \quad \widehat{\mathbf{N}}(x) = \langle \widehat{\mathcal{Q}}^{2m-1} \phi_{\widehat{K}}(x), \phi_{\widehat{K}}(x) \rangle_{\mathcal{H}}^{1/2}.$$

This representation is not fully observable because of the presence of the ideal feature map $\phi_{\overline{K}}$ in the definition of $\widehat{\mathbf{R}}_x$. Nevertheless, it can be used in practice since the representation $\widehat{\mathbf{R}}_x \in L^2_{\mathbf{P}}$ is isometric to the fully observed representation

$$\widehat{\mathcal{R}}_x = \widehat{\mathcal{Q}}^{m-1} \phi_{\widehat{K}}(x), \quad x \in \mathcal{X},$$

in the Hilbert space $(\mathbf{Im}(\mathcal{G}), \|\cdot\|_{\mathcal{G}})$ with the non-observable Hilbert norm $\|u\|_{\mathcal{G}} = \langle \mathcal{G}u, u \rangle^{1/2}$, that could be estimated by $\langle \widehat{\mathcal{Q}}u, u \rangle_{\mathcal{H}}^{1/2}$. For further details we refer to Section 5.2.

In the following we provide non-asymptotic bounds for the error of approximating of the ideal non-normalized representation \mathbf{R}_x with the estimated non-normalized representation $\widehat{\mathbf{R}}_x$.

Proposition 9. *Let χ be defined as in equation (8). For any $x \in \text{supp}(\mathbf{P})$, $f \in L_{\mathbf{P}}^2$, $m > 0$,*

$$\|\mathbf{R}_x - \widehat{\mathbf{R}}_x\|_{L_{\mathbf{P}}^2} \leq \mu(x)^{-1/2} \left((m-1) \|\chi\|_{\infty} \|\widehat{\mathcal{Q}} - \mathcal{G}\|_{\infty} \left(1 + \|\widehat{\mathcal{Q}} - \mathcal{G}\|_{\infty}^{m-2} \right) + \|\chi - 1\|_{\infty} \right)$$

and

$$\left(\int \left\langle \mathbf{R}_x - \widehat{\mathbf{R}}_x, f \right\rangle_{L_{\mathbf{P}}^2}^2 d\mathbf{P}(x) \right)^{1/2} \leq \|f\|_{L_{\mathbf{P}}^2} \left((m-1) \|\chi\|_{\infty} \|\widehat{\mathcal{Q}} - \mathcal{G}\|_{\infty} \left(1 + \|\widehat{\mathcal{Q}} - \mathcal{G}\|_{\infty}^{m-2} \right) + \|\chi - 1\|_{\infty} \right).$$

For the proof we refer to Section 5.2.

The two above propositions link the quality of the approximation (of \overline{K}_{2m} with \widehat{K}_{2m} in Proposition 8 and of \mathbf{R}_x with $\widehat{\mathbf{R}}_x$ in Proposition 9) to the quality of the approximation of the Gram operator \mathcal{G} with $\widehat{\mathcal{Q}}$. In order to qualify the approximation error $\|\widehat{\mathcal{Q}} - \mathcal{G}\|_{\infty}$ we introduce a new intermediate (non completely observable) operator $\overline{\mathcal{G}} : \mathcal{H} \rightarrow \mathcal{H}$ defined as

$$\begin{aligned} \overline{\mathcal{G}}v &= \frac{1}{n} \sum_{i=1}^n \mu(X_i)^{-1} \langle v, \phi_K(X_i) \rangle_{\mathcal{H}} \phi_K(X_i) \\ &= \frac{1}{n} \sum_{i=1}^n \langle v, \phi_{\overline{K}}(X_i) \rangle_{\mathcal{H}} \phi_{\overline{K}}(X_i), \end{aligned}$$

and we observe that the following result holds.

Proposition 10. *Let χ be defined as in equation (8). It holds that*

$$\|\widehat{\mathcal{Q}} - \mathcal{G}\|_{\infty} \leq \|\overline{\mathcal{G}} - \mathcal{G}\|_{\infty} \left(1 + \|\chi^2 - 1\|_{\infty} \right) + \|\chi^2 - 1\|_{\infty}.$$

For the proof we refer to Section 5.3.

Observe that, given $\phi_{\overline{K}}$, the operator $\overline{\mathcal{G}}$ is the empirical version of the Gram operator \mathcal{G} . Moreover, by definition,

$$\chi - 1 = (\mu/\widehat{\mu})^{1/2} - 1$$

where, according to Proposition 2, μ is the quadratic form associated to the Gram operator \mathcal{G}_A and $\hat{\mu}$ is its empirical version. Thus we conclude this section providing a result on the convergence of the empirical Gram estimator to the true one that has been proved in [9] and that appears in [10].

We first introduce some notation. Let \mathcal{H} be a separable Hilbert space and let P be a probability distribution on \mathcal{H} . Let $\mathcal{G} : \mathcal{H} \rightarrow \mathcal{H}$ be the Gram operator

$$\mathcal{G}v = \mathbb{E}_{X \sim P} [\langle v, X \rangle_{\mathcal{H}} X]$$

and consider the empirical estimator

$$\hat{\mathcal{G}}v = \frac{1}{n} \sum_{i=1}^n \langle v, X_i \rangle_{\mathcal{H}} X_i$$

where $X_1, \dots, X_n \in \mathcal{H}$ is an i.i.d. sample drawn according to P . Let $\epsilon > 0$ and let $\sigma > 0$ be a threshold. Let

$$\kappa = \sup_{u \in \mathcal{H}} \frac{\mathbb{E}_{X \sim P} [\langle u, X \rangle_{\mathcal{H}}^4]}{\mathbb{E}_{X \sim P} [\langle u, X \rangle_{\mathcal{H}}^2]^2} < +\infty.$$

Define

$$\begin{aligned} \zeta(t) &= \sqrt{\frac{2.032(\kappa - 1)}{n} \left(\frac{0.73 \operatorname{Tr}(\mathcal{G})}{t} + 4.35 + \log(\epsilon^{-1}) \right)} + \sqrt{\frac{98.5 \kappa \operatorname{Tr}(\mathcal{G})}{nt}} \\ \eta(t) &= \frac{\zeta(\max\{t, \sigma\})}{1 - 4 \zeta(\max\{t, \sigma\})} \\ \tau(t) &= \frac{0.86 \max \|X_i\|_{\mathcal{H}}^4}{n(\kappa - 1) \max\{t, \sigma\}^2} \left(\frac{0.73 \operatorname{Tr}(\mathcal{G})}{\max\{t, \sigma\}} + 4.35 + \log(\epsilon^{-1}) \right), \end{aligned}$$

where $\operatorname{Tr}(\mathcal{G})$ denotes the trace of \mathcal{G} . The following proposition (proved in [10]) holds.

Proposition 11. *Let $\sigma > 0$ be a threshold. With probability at least $1 - 2\epsilon$, for any $u \in \mathcal{H}$, $\|u\|_{\mathcal{H}} = 1$,*

$$\left| \frac{\max \{ \langle \hat{\mathcal{G}}u, u \rangle_{\mathcal{H}}, \sigma \}}{\max \{ \langle \mathcal{G}u, u \rangle_{\mathcal{H}}, \sigma \}} - 1 \right| \leq \eta(\langle \mathcal{G}u, u \rangle_{\mathcal{H}}) + \frac{\tau(\langle \mathcal{G}u, u \rangle_{\mathcal{H}})}{[1 - \tau(\langle \mathcal{G}u, u \rangle_{\mathcal{H}})]_+ [1 - \eta(\langle \mathcal{G}u, u \rangle_{\mathcal{H}})]_+}.$$

As a consequence,

Corollary 12. *With the same notation as before, with probability at least $1 - 2\epsilon$,*

$$\|\mathcal{G} - \widehat{\mathcal{G}}\|_{\infty} \leq \|\mathcal{G}\|_{\infty} \eta(\|\mathcal{G}\|_{\infty}) + \frac{\sigma \tau(\sigma)}{[1 - \tau(\sigma)]_+ [1 - \eta(\sigma)]_+} + \sigma.$$

As explained in [10], the threshold σ can be chosen going to zero as the sample size n goes to infinity. In particular, since $\sigma \tau(\sigma)$ behaves as $\frac{1}{n\sigma^2}$, the optimal value of σ is of order $n^{-1/3}$. As a consequence, according to the above results, we obtain a deviation bound in $n^{-1/3}$ for

$$\sup_{x, y \in \text{supp}(\mathbf{P})} |\widehat{K}_{2m}(x, y) - \overline{K}_{2m}(x, y)|.$$

In order to get a deviation bound in $n^{-1/2}$ we have to use a more robust estimator for the Gram operator \mathcal{G} , defined in [10] (see also [9]), and to split the sample into two parts: the first part is used for the estimation of the kernel \overline{K} and the other one for the construction of the estimator $\widehat{\mathcal{Q}}$.

4 Empirical Results

We present some results showing how the algorithm described in the previous sections simplifies the geometry of the problem and in particular how it groups the points to cluster at the vertices of a simplex. We first provide a toy example on synthetic data and then we test the algorithm in the setting of image analysis.

4.1 A First Example

We consider an i.i.d. set $\{X_1, \dots, X_n\} \subset \mathbb{R}^2$ of $n = 900$ points to cluster, whose configuration is shown in Figure 1 and we fix the maximum number of classes $p = 7$.

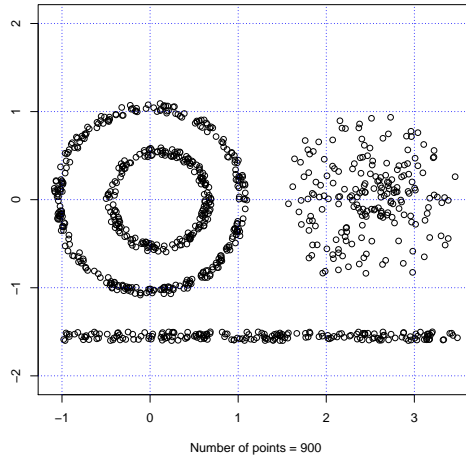


Figure 1: The data configuration.

Figure 2 shows that the new representation, induced by the change of kernel, groups the data points at the vertices of the simplex generated by the largest eigenvectors of the matrix C , defined in equation (7). On the left we plot the projection of the simplex along the two first coordinates. This simple configuration allows us to compute the classification, including the number of clusters, using the straightforward greedy algorithm described in Remark 6.

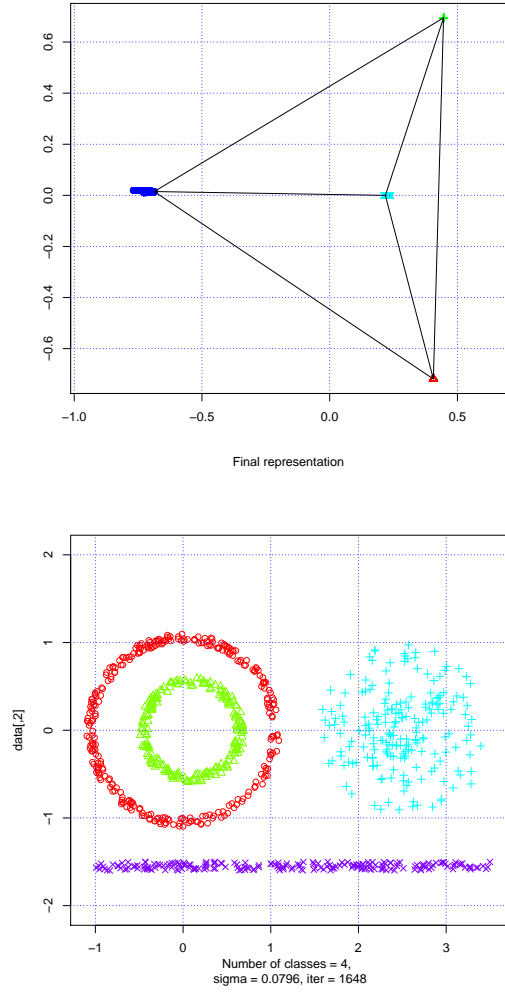


Figure 2: On the left the simplex generated by the eigenvectors of C , on the right classification performed on C .

In Figure 3 we plot the first eigenvalues of M in black (joined by a solid line), the eigenvalues of its iteration M^m in blue (joined by a dashed line) and the eigenvalues of the covariance matrix of the final representation (defined by C) in red (joined by a dash-dotted line). We observe that the first eigenvalues of M are close to one, while there is a remarkable gap between the eigenvalues of its iteration. In particular the size of the gap is larger once we have renormalized it using the matrix C . The number of iterations is automatically estimated and it is equal to 1648.

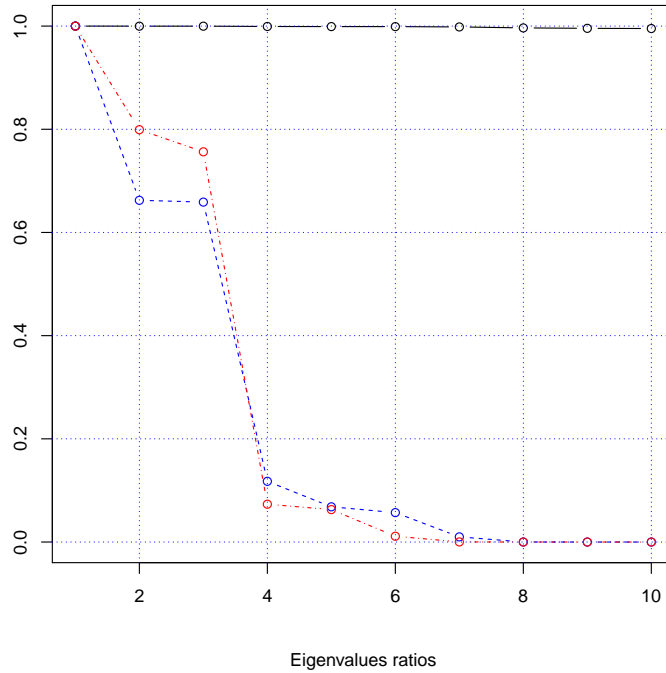


Figure 3: In black (solid line) the eigenvalues of M , in blue (dashed line) those of M^m and in red (dash-dot line) the eigenvalues of C .

4.2 Some Perspectives on Invariant Shape Analysis

As we have already mentioned in the introduction, we now present a small example about image classification showing how our approach may lead to learn transformation-invariant representations from data sets containing

small successive transformations of a same pattern. We briefly describe this approach to get a hint of its potential. We consider two images (Figure 4) and we create our patterns by translating a subwindow of a given size in each image repeatedly, using a translation vector smaller than the subwindow size. In such a way we create a sample consisting of two classes of connected images, as shown in Figure 5. This notion of translation cannot be grasped easily by a mathematical definition, because we do not translate a function but a window of observation. Hence in this case the translation depends on the image content and it may not be easy to model in any realistic situation.

We present, on an example, a successful scenario in which we use first a change of representation in a reproducing kernel Hilbert space to better separate the two classes, and then spectral clustering to shrink each class to a tight blob. This suggests that the so called kernel trick, introduced to better separate classes in the supervised learning framework of support vector machines (SVMs), also works in an unsupervised context. In this setting, we do not separate classes using hyperplanes, since we do not know the class labels which would be necessary to run a SVM, but, instead, we use spectral clustering to finish the work.



Figure 4: The two original images.

Let X_1, \dots, X_n be the sample of images shown in Figure 5. Each photo is represented as a matrix whose entries are the gray values of the corresponding pixels. We apply twice the change of representation described by the change of kernel in order to better separate clusters. We first consider the reproducing kernel Hilbert space \mathcal{H}_1 defined by

$$k_1(x, y) = \exp \left(-\beta_1 \|x - y\|^2 \right)$$



Figure 5: Our sample consisting of two classes of connected images, the first sequence is obtained with a horizontal translation, the second one with a diagonal translation.

and then the reproducing kernel Hilbert space \mathcal{H}_2 defined by

$$\begin{aligned} k_2(x, y) &= \exp\left(-\beta_2\|x - y\|_{\mathcal{H}_1}^2\right) = \exp\left(-2\beta_2(1 - k_1(x, y))\right) \\ &= \exp\left[-2\beta_2\left(1 - \exp\left(-\beta_1\|x - y\|^2\right)\right)\right] \end{aligned}$$

where $\beta_1, \beta_2 > 0$ are obtained as described in Section 2.3. Define the new kernel

$$K(x, y) = \exp\left(-\beta\|x - y\|_{\mathcal{H}_2}^2\right),$$

where the parameter $\beta > 0$ is chosen again as in Section 2.3, and apply the algorithm described in Section 3.2.

In Figure 6 we compare the representation of the images in the initial space and in the space \mathcal{H}_2 . On the left we present the projection of the sample onto the space spanned by the first two largest eigenvectors of the matrix of inner products between images $\langle X_i, X_j \rangle$. On the right we plot the projection onto the space spanned by the two largest eigenvectors of the matrix of inner products $k_2(X_i, X_j)$ in \mathcal{H}_2 . We observe that in the first representation the two classes intersect each other while in the second one, after the change of representation, they are already separated.

To conclude, Figure 7 shows the final representation. Here the data points are projected onto the space spanned by the two largest eigenvectors of the matrix M^m . In this case the number of iteration m is of order of 30.000.

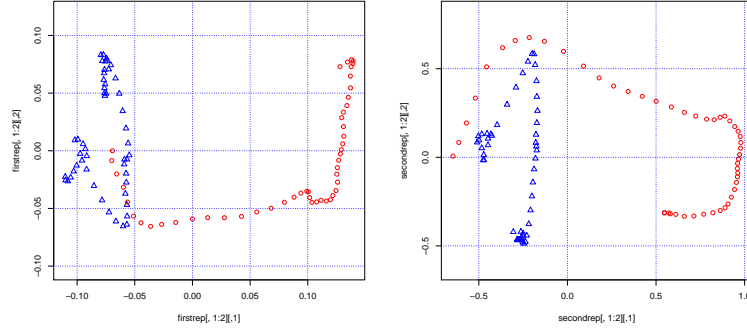


Figure 6: On the left the projection onto the space spanned by the two largest eigenvectors of $\langle X_i, X_j \rangle$, on the right the projection onto the space spanned by the two largest eigenvectors of $k_2(X_i, X_j)$.

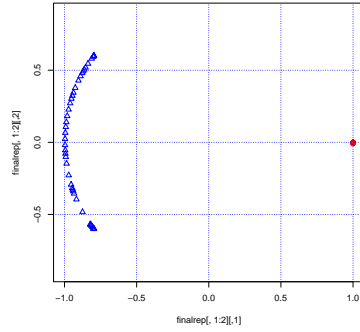


Figure 7: The projection onto the space spanned by the two largest eigenvectors of M^m .

5 Proofs

We introduce a technical result that is useful in the following. Let \mathcal{G} be the Gram operator defined in equation (4) such that $\|\mathcal{G}\|_\infty = 1$ and let $\hat{\mathcal{Q}}$ be its estimator introduced in Section 3.1.

Lemma 13. *For any $m > 0$*

$$\begin{aligned}\|\hat{\mathcal{Q}}^m - \mathcal{G}^m\|_\infty &\leq \left(1 + \|\hat{\mathcal{Q}} - \mathcal{G}\|_\infty\right)^m - 1 \\ &\leq m\|\hat{\mathcal{Q}} - \mathcal{G}\|_\infty \left(1 + \|\hat{\mathcal{Q}} - \mathcal{G}\|_\infty\right)^{m-1}.\end{aligned}$$

Proof. Since

$$\hat{\mathcal{Q}}^m - \mathcal{G}^m = \sum_{k=0}^{m-1} \hat{\mathcal{Q}}^k (\hat{\mathcal{Q}} - \mathcal{G}) \mathcal{G}^{m-k-1},$$

using the fact that $\|\mathcal{G}\|_\infty = 1$ we get

$$\|\hat{\mathcal{Q}}^m - \mathcal{G}^m\|_\infty \leq \|\hat{\mathcal{Q}} - \mathcal{G}\|_\infty \sum_{k=0}^{m-1} \|\hat{\mathcal{Q}}\|_\infty^k.$$

Hence, as $\|\hat{\mathcal{Q}}\|_\infty \leq 1 + \|\hat{\mathcal{Q}} - \mathcal{G}\|_\infty$, we conclude

$$\begin{aligned}\|\hat{\mathcal{Q}}^m - \mathcal{G}^m\|_\infty &\leq \left(1 + \|\hat{\mathcal{Q}} - \mathcal{G}\|_\infty\right)^m - 1 \\ &= \|\hat{\mathcal{Q}} - \mathcal{G}\|_\infty \sum_{k=0}^{m-1} \left(1 + \|\hat{\mathcal{Q}} - \mathcal{G}\|_\infty\right)^k \leq m\|\hat{\mathcal{Q}} - \mathcal{G}\|_\infty \left(1 + \|\hat{\mathcal{Q}} - \mathcal{G}\|_\infty\right)^{m-1}.\end{aligned}$$

■

□

We also introduce the intermediate operator $\tilde{\mathcal{G}} : \mathcal{H} \rightarrow \mathcal{H}$

$$\tilde{\mathcal{G}}u = \int \langle u, \phi_{\hat{K}}(z) \rangle_{\mathcal{H}} \phi_{\hat{K}}(z) \mathrm{dP}(z)$$

where, according to the notation of Section 3, $\phi_{\hat{K}}(x) = \chi(x)\phi_{\bar{K}}(x)$ and $\chi(x) = (\mu(x)/\hat{\mu}(x))^{1/2}$.

5.1 Proof of Proposition 8

To prove the first inequality, we define

$$\mathcal{E}(x, y) = |\widehat{K}_{2m}(x, y) - \overline{K}_{2m}(x, y)|$$

and we observe that, according to Lemma 13, it is sufficient to show that

$$\mathcal{E}(x, y) \leq \frac{\max\{1, \|\chi\|_\infty\}^2}{\mu(x)^{1/2}\mu(y)^{1/2}} \left(\|\widehat{\mathcal{Q}}^{2m-1} - \mathcal{G}^{2m-1}\|_\infty + 2\|\chi - 1\|_\infty \right). \quad (9)$$

By definition,

$$\begin{aligned} \mathcal{E}(x, y) &= \left| \left\langle \widehat{\mathcal{Q}}^{2m-1} \phi_{\widehat{K}}(x), \phi_{\widehat{K}}(y) \right\rangle_{\mathcal{H}} - \left\langle \mathcal{G}^{2m-1} \phi_{\overline{K}}(x), \phi_{\overline{K}}(y) \right\rangle_{\mathcal{H}} \right| \\ &\leq \left| \left\langle (\widehat{\mathcal{Q}}^{2m-1} - \mathcal{G}^{2m-1}) \phi_{\widehat{K}}(x), \phi_{\widehat{K}}(y) \right\rangle_{\mathcal{H}} \right| + \left| \left\langle \mathcal{G}^{2m-1} (\phi_{\widehat{K}}(x) - \phi_{\overline{K}}(x)), \phi_{\widehat{K}}(y) \right\rangle_{\mathcal{H}} \right| \\ &\quad + \left| \left\langle \mathcal{G}^{2m-1} \phi_{\overline{K}}(x), (\phi_{\widehat{K}}(y) - \phi_{\overline{K}}(y)) \right\rangle_{\mathcal{H}} \right|. \end{aligned}$$

Recalling the definition of $\phi_{\widehat{K}}$, we get

$$\mathcal{E}(x, y) \leq \|\widehat{\mathcal{Q}}^{2m-1} - \mathcal{G}^{2m-1}\|_\infty \|\chi\|_\infty^2 \|\phi_{\overline{K}}(x)\|_{\mathcal{H}} \|\phi_{\overline{K}}(y)\|_{\mathcal{H}} + \|\chi - 1\|_\infty (\|\chi\|_\infty + 1) \|\phi_{\overline{K}}(x)\|_{\mathcal{H}} \|\phi_{\overline{K}}(y)\|_{\mathcal{H}},$$

where, since $K(x, x) = 1$,

$$\|\phi_{\overline{K}}(x)\|_{\mathcal{H}}^2 = \overline{K}(x, x) = \frac{K(x, x)}{\mu(x)^{1/2}\mu(x)^{1/2}} = \frac{1}{\mu(x)}.$$

This proves equation (9). To prove the second bound we define

$$\mathcal{E}(x) = \left(\int \mathcal{E}(x, y)^2 \mathrm{dP}(y) \right)^{1/2}$$

so that using Lemma 13 again, it is sufficient to show that

$$\mathcal{E}(x) \leq \frac{\max\{1, \|\chi\|_\infty\}^2}{\mu(x)^{1/2}} \left(\|\widehat{\mathcal{Q}}^{2m-1} - \mathcal{G}^{2m-1}\|_\infty + 2\|\chi - 1\|_\infty \right).$$

Observe that

$$\begin{aligned}
\mathcal{E}(x) &= \left(\int \left(\langle \widehat{\mathcal{Q}}^{2m-1} \phi_{\widehat{K}}(x), \phi_{\widehat{K}}(y) \rangle_{\mathcal{H}} - \langle \mathcal{G}^{2m-1} \phi_{\overline{K}}(x), \phi_{\overline{K}}(y) \rangle_{\mathcal{H}} \right)^2 dP(y) \right)^{1/2} \\
&\leq \left(\int \langle (\widehat{\mathcal{Q}}^{2m-1} - \mathcal{G}^{2m-1}) \phi_{\widehat{K}}(x), \phi_{\widehat{K}}(y) \rangle_{\mathcal{H}}^2 dP(y) \right)^{1/2} \\
&\quad + \left(\int \langle \mathcal{G}^{2m-1} (\phi_{\widehat{K}}(x) - \phi_{\overline{K}}(x)), \phi_{\widehat{K}}(y) \rangle_{\mathcal{H}}^2 dP(y) \right)^{1/2} \\
&\quad + \left(\int \langle \mathcal{G}^{2m-1} \phi_{\overline{K}}(x), \phi_{\widehat{K}}(y) - \phi_{\overline{K}}(y) \rangle_{\mathcal{H}}^2 dP(y) \right)^{1/2}.
\end{aligned}$$

By definition of $\widetilde{\mathcal{G}}$ we get

$$\begin{aligned}
\mathcal{E}(x) &\leq \|\widetilde{\mathcal{G}}^{1/2}(\widehat{\mathcal{Q}}^{2m-1} - \mathcal{G}^{2m-1})\phi_{\widehat{K}}(x)\|_{\mathcal{H}} + \|\widetilde{\mathcal{G}}^{1/2}\mathcal{G}^{2m-1}(\phi_{\widehat{K}}(x) - \phi(x))\|_{\mathcal{H}} \\
&\quad + \|\chi - 1\|_{\infty} \|\mathcal{G}^{2m-1/2}\phi(x)\|_{\mathcal{H}}.
\end{aligned}$$

Thus using the fact that $\|\mathcal{G}\|_{\infty} = 1$ and that, for any $u \in \mathcal{H}$,

$$\|\widetilde{\mathcal{G}}^{1/2}u\|_{\mathcal{H}}^2 = \langle \widetilde{\mathcal{G}}u, u \rangle_{\mathcal{H}} = \int \langle u, \phi_{\widehat{K}}(y) \rangle_{\mathcal{H}}^2 dP(y) \leq \|\chi\|_{\infty}^2 \langle \mathcal{G}u, u \rangle,$$

we conclude.

5.2 Proof of Proposition 9

In order to prove the two inequalities we need to introduce some preliminary results. Consider the operator $\mathcal{S} : \mathcal{H} \rightarrow L_{\mathbb{P}}^2$

$$\mathcal{S}(u) : x \mapsto \langle u, \phi_{\overline{K}}(x) \rangle_{\mathcal{H}}.$$

Introduce the operator $\mathbf{G} : L_{\mathbb{P}}^2 \rightarrow L_{\mathbb{P}}^2$

$$\mathbf{G}(f)(x) = \mathcal{S}\mathcal{S}^*f(x) = \int \overline{K}(x, y) f(y) dP(y)$$

and observe that the Gram operator \mathcal{G} rewrites as

$$\mathcal{G}u = \mathcal{S}^*\mathcal{S}u = \int \langle u, \phi_{\overline{K}}(z) \rangle_{\mathcal{H}} \phi_{\overline{K}}(z) dP(z).$$

Moreover $\|\mathbf{G}\|_{\infty} = \|\mathcal{G}\|_{\infty} = 1$.

Lemma 14. *We have*

$$\mathbf{Im}(\mathbf{G}) = \mathcal{S}(\mathbf{Im}(\mathcal{G})) = \mathcal{S}(\mathbf{Im}(\tilde{\mathcal{G}})).$$

Proof. We observe that, since \mathbf{G} is symmetric, $\mathbf{Im}(\mathbf{G}) = \mathbf{Im}(\mathbf{G}^2)$, where $\mathbf{G}^2 = \mathcal{S}\mathcal{G}\mathcal{S}^*$. Thus $\mathbf{Im}(\mathbf{G}) = \mathbf{Im}(\mathbf{G}^2) \subset \mathcal{S}(\mathbf{Im}(\mathcal{G}))$. Since $\mathcal{S}\mathcal{G} = \mathbf{G}\mathcal{S}$, we conclude that $\mathcal{S}(\mathbf{Im}(\mathcal{G})) \subset \mathbf{Im} \mathbf{G}$.

To prove the second identity, we remark that, since $\chi(x) > 0$,

$$\begin{aligned} \ker(\tilde{\mathcal{G}}) &= \left\{ u \in \mathcal{H}, \int \langle u, \phi_{\hat{K}}(x) \rangle^2 dP(x) = 0 \right\} \\ &= \left\{ u \in \mathcal{H}, \int \langle u, \phi(x) \rangle^2 dP(x) = 0 \right\} = \ker(\mathcal{G}). \end{aligned}$$

Consequently, since $\mathbf{Im}(\mathcal{G}) = \ker(\mathcal{G})^\perp$, we get

$$\overline{\text{span}}(\phi(\text{supp}(P))) = \ker(\mathcal{G})^\perp = \mathbf{Im}(\mathcal{G}) = \mathbf{Im}(\tilde{\mathcal{G}}) = \overline{\text{span}}(\phi_{\hat{K}}(\text{supp}(P))).$$

■

□

Therefore any $f \in \mathbf{Im}(\mathbf{G})$ is of the form $f = \mathcal{S}u$, with $u \in \mathbf{Im}(\mathcal{G})$, so that we can estimate

$$\|f\|_{L_{\mathbb{P}}^2}^2 = \langle \mathcal{G}u, u \rangle_{\mathcal{H}}$$

by $\langle \hat{\mathcal{Q}}u, u \rangle_{\mathcal{H}}$. The estimation error is bounded as described in the following lemma.

Lemma 15. *For any $u, v \in \mathcal{H}$,*

$$|\langle \mathcal{S}u, \mathcal{S}v \rangle_{L_{\mathbb{P}}^2} - \langle \hat{\mathcal{Q}}u, v \rangle_{\mathcal{H}}| \leq \|\mathcal{G} - \hat{\mathcal{Q}}\|_{\infty} \|u\|_{\mathcal{H}} \|v\|_{\mathcal{H}}.$$

In particular,

$$|\|\mathcal{S}u\|_{L_{\mathbb{P}}^2}^2 - \langle \hat{\mathcal{Q}}u, u \rangle_{\mathcal{H}}| \leq \|\mathcal{G} - \hat{\mathcal{Q}}\|_{\infty} \|u\|_{\mathcal{H}}^2.$$

Proof. It is sufficient to observe that $\langle \mathcal{S}u, \mathcal{S}v \rangle_{L_{\mathbb{P}}^2} = \langle \mathcal{G}u, v \rangle_{\mathcal{H}}$. ■

□

We now observe that

$$\begin{aligned}\mathbf{R}_x &= \mathcal{S}\mathcal{G}^{m-1}\phi_{\overline{K}}(x) \in \mathbf{Im}(\mathbf{G}) \\ \widehat{\mathbf{R}}_x &= \mathcal{S}\widehat{\mathcal{Q}}^{m-1}\phi_{\widehat{K}}(x) \in \mathbf{Im}(\mathbf{G}) \quad \text{almost surely.}\end{aligned}\tag{10}$$

Indeed, by definition,

$$\widehat{\mathbf{R}}_x = \mathcal{S}\widehat{\mathcal{Q}}^{m-1}\phi_{\widehat{K}}(x) \quad \text{and} \quad \mathbf{R}_x = \mathcal{S}\mathcal{G}^{m-1}\phi_{\overline{K}}(x).$$

Hence, since

$$\mathbf{Im}(\mathbf{G}) = \mathcal{S}(\mathbf{Im}(\mathcal{G})) = \mathcal{S}(\overline{\text{span}} \phi_{\overline{K}}(\text{supp}(\mathbf{P}))),$$

we conclude that $\mathbf{R}_x \in \mathbf{Im}(\mathbf{G})$. Moreover, since

$$\begin{aligned}\text{span}\{\phi_{\widehat{K}}(X_n), \dots, \phi_{\widehat{K}}(X_{2n})\} &= \text{span}\{\phi_{\overline{K}}(X_n), \dots, \phi_{\overline{K}}(X_{2n})\} \\ &\subset \overline{\text{span}}(\phi_{\overline{K}}(\text{supp}(\mathbf{P}))) \quad \text{almost surely,}\end{aligned}$$

it is also true that $\widehat{\mathbf{R}}_x \in \mathbf{Im}(\mathbf{G})$.

We can now prove the two bounds presented in the proposition. Define

$$\begin{aligned}\mathcal{E}_r(x) &= \|\mathbf{R}_x - \widehat{\mathbf{R}}_x\|_{L_{\mathbf{P}}^2} \\ \mathcal{E}_c(f) &= \left(\int \langle \mathbf{R}_x - \widehat{\mathbf{R}}_x, f \rangle_{L_{\mathbf{P}}^2}^2 d\mathbf{P}(x) \right)^{1/2}.\end{aligned}$$

According to Lemma 13, it is sufficient to show that, for any $x \in \text{supp}(\mathbf{P})$, $f \in L_{\mathbf{P}}^2$,

$$\begin{aligned}\mathcal{E}_r(x) &\leq \mu(x)^{-1/2} \left(\|\chi\|_{\infty} \|\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1}\|_{\infty} + \|\chi - 1\|_{\infty} \right) \\ \mathcal{E}_c(f) &\leq \|f\|_{L_{\mathbf{P}}^2} \left(\|\chi\|_{\infty} \|\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1}\|_{\infty} + \|\chi - 1\|_{\infty} \right).\end{aligned}$$

In order to prove the first inequality, we observe that

$$\begin{aligned}\mathcal{E}_r(x) &= \|\mathcal{S}\widehat{\mathcal{Q}}^{m-1}\phi_{\widehat{K}}(x) - \mathcal{S}\mathcal{G}^{m-1}\phi_{\overline{K}}(x)\|_{L_{\mathbf{P}}^2} \\ &\leq \|\mathcal{S}(\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1})\phi_{\widehat{K}}(x)\|_{L_{\mathbf{P}}^2} + \|\mathcal{S}\mathcal{G}^{m-1}(\phi_{\widehat{K}}(x) - \phi_{\overline{K}}(x))\|_{L_{\mathbf{P}}^2}.\end{aligned}$$

Since $\|\mathcal{S}u\|_{L_{\mathbf{P}}^2}^2 = \langle \mathcal{S}^* \mathcal{S}u, u \rangle_{\mathcal{H}} = \langle \mathcal{G}u, u \rangle_{\mathcal{H}}$, then $\|\mathcal{S}\|_{\infty} = \|\mathcal{G}\|_{\infty}^{1/2} = 1$ and hence, recalling the definition of $\phi_{\widehat{K}}$, we get

$$\mathcal{E}_r(x) \leq \|\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1}\|_{\infty} \|\chi\|_{\infty} \|\phi_{\overline{K}}(x)\|_{\mathcal{H}} + \|\chi - 1\|_{\infty} \|\phi_{\overline{K}}(x)\|_{\mathcal{H}}.$$

We now prove the second bound. Let $\mathbf{\Pi} : L_{\mathbb{P}}^2 \rightarrow L_{\mathbb{P}}^2$ be the orthogonal projector on $\mathbf{Im}(\mathbf{G})$. Since, according to equation (10), almost surely, $\widehat{\mathbf{R}}_x - \mathbf{R}_x \in \mathbf{Im}(\mathbf{G})$, for any $x \in \mathcal{X}$, then

$$\left\langle \widehat{\mathbf{R}}_x - \mathbf{R}_x, f \right\rangle_{L_{\mathbb{P}}^2} = \left\langle \widehat{\mathbf{R}}_x - \mathbf{R}_x, \mathbf{\Pi}(f) \right\rangle_{L_{\mathbb{P}}^2} \quad \text{almost surely.}$$

Moreover, since $\mathbf{Im}(\mathbf{G}) = \mathcal{S}(\mathbf{Im}(\mathcal{G}))$, there is $u \in \mathbf{Im}(\mathcal{G})$ such that $\mathbf{\Pi}(f) = \mathcal{S}u$. We can then write

$$\begin{aligned} \left\langle \widehat{\mathbf{R}}_x - \mathbf{R}_x, f \right\rangle_{L_{\mathbb{P}}^2} &= \left\langle \widehat{\mathbf{R}}_x - \mathbf{R}_x, \mathcal{S}u \right\rangle_{L_{\mathbb{P}}^2} \\ &= \left\langle \mathcal{S}(\widehat{\mathcal{Q}}^{m-1}\phi_{\widehat{K}}(x) - \mathcal{G}^{m-1}\phi_{\overline{K}}(x)), \mathcal{S}u \right\rangle_{L_{\mathbb{P}}^2} \\ &= \left\langle \widehat{\mathcal{Q}}^{m-1}\phi_{\widehat{K}}(x) - \mathcal{G}^{m-1}\phi_{\overline{K}}(x), \mathcal{G}u \right\rangle_{\mathcal{H}} \\ &= \left\langle (\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1})\phi_{\widehat{K}}(x), \mathcal{G}u \right\rangle_{\mathcal{H}} + \left\langle \mathcal{G}^{m-1}(\phi_{\widehat{K}}(x) - \phi_{\overline{K}}(x)), \mathcal{G}u \right\rangle_{\mathcal{H}}. \end{aligned}$$

Therefore, similarly as before, we get

$$\begin{aligned} \mathcal{E}_c(f) &\leq \|\widehat{\mathcal{G}}^{1/2}(\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1})\mathcal{G}u\|_{\mathcal{H}} + \|\chi - 1\|_{\infty} \left(\int \langle \mathcal{G}^{m-1}\phi_{\overline{K}}(x), \mathcal{G}u \rangle_{\mathcal{H}}^2 d\mathbb{P}(x) \right)^{1/2} \\ &= \|\widehat{\mathcal{G}}^{1/2}(\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1})\mathcal{G}u\|_{\mathcal{H}} + \|\chi - 1\|_{\infty} \|\mathcal{G}^{m+1/2}u\|_{\mathcal{H}} \\ &\leq \|\chi\|_{\infty} \|\widehat{\mathcal{Q}}^{m-1} - \mathcal{G}^{m-1}\|_{\infty} \|\mathcal{G}^{1/2}u\|_{\mathcal{H}} + \|\chi - 1\|_{\infty} \|\mathcal{G}^{1/2}u\|_{\mathcal{H}}. \end{aligned}$$

We conclude observing that

$$\|\mathcal{G}^{1/2}u\|_{\mathcal{H}} = \langle \mathcal{S}u, \mathcal{S}u \rangle_{L_{\mathbb{P}}^2} = \|\mathbf{\Pi}(f)\|_{L_{\mathbb{P}}^2} \leq \|f\|_{L_{\mathbb{P}}^2}.$$

5.3 Proof of Proposition 10

Observe that

$$\|\widehat{\mathcal{Q}} - \mathcal{G}\|_{\infty} \leq \|\widehat{\mathcal{Q}} - \overline{\mathcal{G}}\|_{\infty} + \|\overline{\mathcal{G}} - \mathcal{G}\|_{\infty}.$$

Moreover, for any $u \in \mathcal{H}$, such that $\|u\|_{\mathcal{H}} = 1$, recalling that $\phi_{\overline{K}}(x) = \mu(x)^{-1/2}\phi_K(x)$,

$$\begin{aligned} \langle \widehat{\mathcal{Q}}u, u \rangle_{\mathcal{H}} - \langle \overline{\mathcal{G}}u, u \rangle_{\mathcal{H}} &= \frac{1}{n} \sum_{i=1}^n \left(\widehat{\mu}(X_i)^{-1} - \mu(X_i)^{-1} \right) \langle u, \phi_K(X_i) \rangle_{\mathcal{H}}^2 \\ &= \frac{1}{n} \sum_{i=1}^n \mu(X_i)^{-1} \left(\chi(X_i)^2 - 1 \right) \langle u, \phi_K(X_i) \rangle_{\mathcal{H}}^2. \end{aligned}$$

Thus

$$\begin{aligned}\|\hat{\mathcal{Q}} - \bar{\mathcal{G}}\|_{\infty} &\leq \|\chi^2 - 1\|_{\infty} \sup_{\|u\|_{\mathcal{H}}=1} \frac{1}{n} \sum_{i=1}^n \mu(X_i)^{-1} \langle u, \phi_K(X_i) \rangle_{\mathcal{H}}^2 \\ &= \|\chi^2 - 1\|_{\infty} \sup_{\|u\|_{\mathcal{H}}=1} \frac{1}{n} \sum_{i=1}^n \langle u, \phi_{\bar{K}}(X_i) \rangle_{\mathcal{H}}^2 = \|\chi^2 - 1\|_{\infty} \|\bar{\mathcal{G}}\|_{\infty}.\end{aligned}$$

Using the fact that $\|\mathcal{G}\|_{\infty} = 1$ we conclude that

$$\|\hat{\mathcal{Q}} - \bar{\mathcal{G}}\|_{\infty} \leq \|\chi^2 - 1\|_{\infty} \left(1 + \|\bar{\mathcal{G}} - \mathcal{G}\|_{\infty}\right),$$

which proves the proposition.

Bibliography

- [1] Joakim Andén and Stéphane Mallat. Multiscale scattering for audio classification. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 657–662, Miami (Florida), USA, October 24–28 2011.
- [2] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, June 2003.
- [3] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1872–1886, August 2013.
- [4] Olivier Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.
- [5] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.
- [6] Nello Cristianini, John Shawe-Taylor, and Jaz S. Kandola. Spectral kernel methods for clustering. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3–8, 2001, Vancouver, British Columbia, Canada]*, pages 649–655, 2001.
- [7] W. E. Donath and A. J. Hoffman. Lower bounds for the partitioning of graphs. *IBM J. Res. Develop.*, 17:420–425, 1973.
- [8] Miroslav Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Math. J.*, 25(100)(4):619–633, 1975.
- [9] Ilaria Giulini. Generalization bounds for random samples in hilbert spaces. *PhD thesis*.
- [10] Ilaria Giulini. Robust dimension-free gram operator estimates. *preprint arXiv:1511.06259*.

- [11] Michael I. Jordan and Francis R. Bach. Learning spectral clustering. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2003.
- [12] Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007.
- [13] Stéphane Mallat. Group invariant scattering. *Comm. Pure Appl. Math.*, 65(10):1331–1398, 2012.
- [14] Marina Meila and Jianbo Shi. Learning segmentation by random walks. In *Advances in Neural Information Processing Systems*, pages 873–879. MIT Press, 2001.
- [15] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press, 2001.
- [16] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, August 2000.
- [17] Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *Ann. Statist.*, 36(2):555–586, 2008.